The effect of rational selection of training sets from an imbalanced AhR activation dataset on QSAR models accuracy and applicability domain coverage for a large set of REACH substances

Klimenko, Kyrylo Oleksandrovych; Abildgaard Rosenberg, Sine; Dybdahl, Marianne; Wedebye, Eva Bay; Nikolov, Nikolai Georgiev

Publication date:
2018

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available

The effect of rational selection of training sets from an imbalanced AhR activation dataset on QSAR models accuracy and applicability domain coverage for a large set of REACH substances

Klimenko Kyrylo,[1,*] Rosenberg Sine,[1] Dybdahl Marianne,[1] Wedebye Eva ,[1] and Nikolov Nikolai[1]

[1]*National Food Institute, Technical University of Denmark, Kemitorvet, Building 201, 2800, Kgs. Lyngby, Denmark*

The aryl hydrocarbon receptor (AhR) is a ligand-dependent transcription factor that regulates the expression of multiple genes of importance for among other things organ development, the immune system and the metabolism of exogenous and endogenous small molecules. AhR activation by industrial chemical substances may lead to increased turnover of the endogenous estrogen and thyroid hormones, possibly resulting in adverse outcomes.

A PubChem experimental data set on AhR activation with 324,858 chemical substances which is heavily skewed towards inactives was used to develop QSAR models using a stepwise rational training set selection approach. After randomly selecting equal proportions of actives and inactives to make initial models, predictions of large external inactive selection sets were made and used to rationally select and add inactives to the training sets. This was done in an iterative process to produce final models. Two approaches were taken to select additional training set compounds: in the first approach substances were added that were either predicted incorrectly as positives or were out of structural or probability applicability domain, and in the second approach substances were added with a more focused scope to optimize the applicability domain for REACH substances. Final models resulting from both approaches were used to predict approximately 80,000 REACH industrial chemical substances. The advantages and applicability of each approach to predicting potential endocrine disruptors are discussed.

Section

	(Q)SAR models for regulatory use
X	Models for human health effects
	Models for ecotoxicological and environmental effects
X	Protein-ligand interactions, *in silico* studies related to toxicological effects
	Software and tools

Presentation

X	oral
	poster